

Prepare a dataset to compare Species Sensitivity to Toxic Substances

Richard Meitern

2023-12-18

Introduction

The aim of this document is to prepare a dataset for comparing the sensitivity of species to toxic substances by order. More specifically, we want to know which orders are most sensitive to which toxic substances, if the species sensitivity within the order differs depending on species life history traits and which chemicals or chemical classes are most toxic to which orders or species. In order to do this we will plot out the 50% lethal concentration (LC) of species and orders to toxic substances.

Import and Prepare Data

There are essentially three data sources:

- `species_LH` contains the species information like name, order, family, etc.
- `ecotox_name_VS_compound` contains the names of the toxic substances
- `toxData` contains the toxicity data for each species and toxic substance

The import loads the sources and combines them into a single data frame `df`.

Unify Column Contents in Toxicity Data

The data in the species toxicity data (`toxdata`) is not uniform. For example, the column “`conc_units`” contains the values: AI mg/L, ul/L, mmol/kg, ml/L, AI mg/kg bdwt, ml, um, NA, %, mg/fish, mg, ug/egg. The following code unifies the values in the columns and selects for each column the most common value.

```
## The toxicity data contains 4603 (65.89%) usable rows.
```

We looked through the information in the toxicity table and extracted the comparable rows. The usable rows are those that have a *50% LC* value for the endpoint, the measured value was *Mortality*, the units are converted to *mg/mL*, the organisms were adults and the exposure was done in a lab environment by exposing the chemical by dosing it into the water. If the values for dosing, organism lifestage or exposure environment was missing the data was kept, effectively assuming that the missing values were the same as the most common or expected value (eg the exposure was assumed to be done in a lab environment as for most studies this data was not reported but could be assumed to be the case).

```
## Originally the mean concentration for LC is missing for: 207 rows!
```

```
## After imputing from the min max-mean, mean concentration for LC is missing for: 1 rows!
```

Fix Species Names

The species names in the species table and the toxicity table are not always the same. The following code fixes this problem by replacing the names in the toxicity table with unified names in the species table.

Nevertheless, there are still some species names in the `toxData` missing in the species table.

```
## Species missing in the species table present in the tox table make up 33 (0.7%) rows of the tox table
```

```
## Priopidichthys sp.  
## Mugil sp.  
## Cyprinidae  
## Boleophthalmus sp.  
## Tilapia sp.  
## Perca sp.  
## Pleuronectiformes  
## Barbus sp.  
## Synodontis sp.  
## Gobius sp.  
## Centrarchidae  
## Osteichthyes
```

These are genus or order level species names, we will remove them from the dataset with the merge.

```
## Species missing in the tox table present in the species table:
```

```
## c('Channa marulius' = '',  
## 'Clarias gariepinus' = '',  
## 'Macropodus opercularis' = '',  
## 'Pseudosphromenus cupanus' = '',  
## 'Acipenser transmontanus' = '',  
## 'Acrossocheilus paradoxus' = '',  
## 'Aldrichetta forsteri' = '',  
## 'Ambloplites rupestris' = '',  
## 'Apeltes quadracus' = '',  
## 'Aphanius dispar' = '',  
## 'Aplocheilus melanostigma' = '',  
## 'Atherinops affinis' = '',  
## 'Atherinosoma microstoma' = '',  
## 'Barbus javanicus' = '',  
## 'Brevoortia tyrannus' = '',  
## 'Candidia barbatus' = '',  
## 'Capoeta fusca' = '',  
## 'Chanodichthys erythropterus' = '',  
## 'Chanos chanos' = '',  
## 'Chirostoma jordani' = '',  
## 'Chrysophrys major' = '',  
## 'Coregonus fera' = '',  
## 'Coryphaena hippurus' = '',  
## 'Cottus bairdi' = '',  
## 'Cottus cognatus' = '',  
## 'Cynoglossus joyneri' = '',  
## 'Dicentrarchus labrax' = '',
```

```

## 'Engraulis japonicus' = '',
## 'Engraulis mordax' = '',
## 'Etheostoma rubrum' = '',
## 'Etheostoma spectabile' = '',
## 'Fundulus grandis' = '',
## 'Halobatrachus didactylus' = '',
## 'Hypophthalmichthys molitrix' = '',
## 'Hypophthalmichthys nobilis' = '',
## 'Lampetra tridentata' = '',
## 'Lepomis humilis' = '',
## 'Liza ramado' = '',
## 'Liza vaigiensis' = '',
## 'Lutjanus argentimaculatus' = '',
## 'Melanogrammus aeglefinus' = '',
## 'Melanotaenia fluviatilis' = '',
## 'Micropterus dolomieu' = '',
## 'Mugil saliens' = '',
## 'Mylopharyngodon piceus' = '',
## 'Nothobranchius furzeri' = '',
## 'Nothobranchius guentheri' = '',
## 'Odontesthes regia' = '',
## 'Onychostoma barbata' = '',
## 'Opsanus beta' = '',
## 'Oryzias melastigma' = '',
## 'Petromyzon marinus' = '',
## 'Piaractus mesopotamicus' = '',
## 'Plagiognathops microlepis' = '',
## 'Pleuronectes platessa' = '',
## 'Pomatoschistus microps' = '',
## 'Prosopium williamsoni' = '',
## 'Psetta maxima' = '',
## 'Pseudaphritis urvillii' = '',
## 'Pseudopleuronectes yokohamae' = '',
## 'Ptychocheilus oregonensis' = '',
## 'Rhamdia quelen' = '',
## 'Rutilus kutum' = '',
## 'Scorpaenichthys marmoratus' = '',
## 'Sparus aurata' = '',
## 'Tilapia guineensis')

## Warning: There are still missing taxon ids. These are:
## c("Microparchax johnstoni", "Coregonus fera", "Puntius stigma")
## Coregonus fera is extinct Salmoniform not in NCBI
## Puntius stigma is not found in NCBI maybe Leuciscus stigma?
## Microparchax johnstoni is not found in NCBI

```

Prepare the Dataset Columns for Analysis

We are going to do some data processing to make the data more suitable for analysis. First we convert the characters to factors.

Then, to compare the toxicity of different compounds within a class of chemicals we need to standardize the toxicity values. We do this by dividing the individual species LC50 with the mean LC50 value for all the species combined. This gives a relative toxicity value for each for each compound for a given species and

enables to compare compounds with a really different LC50 concentrations. The idea behind this is that the toxicity of a compound to biological organisms starts generally at a threshold and this is really different for different compounds but not so different for different organisms.

```
## Warning in ifelse(df$solub_water_mg_l == "insoluble", 0,
## as.numeric(df$solub_water_mg_l)): NAs introduced by coercion

## Warning in ifelse(df$log_k_ow == "insoluble", 0, as.numeric(df$log_k_ow)): NAs
## introduced by coercion
```

Substance	Mean LC50	Median LC50	Nr Obs
Trifluralin (2,6-Dinitro-N,N-dipropyl-4-...	3.9	0.13	126
Chlorpyrifos (O,O-Diethyl O-(3,5,6-trich...	0.56	0.073	307
Pentachlorophenol	0.27	0.2	571
4-nonylphenol	0.35	0.3	285
Cd and its compounds	35	3.1	904
Endosulfan (6,7,8,9,10,10-Hexachloro-1,5...	0.16	0.003	479
Tributyltin chloride	0.012	0.0073	20
Hexachlorocyclohexane	79	0.15	507
Atrazine (6-Chloro-N2-ethyl-N4-(propan-2...	30	20	92
Simazine (6-Chloro-N,N'-diethyl-1,3,5-tr...	143	49	73
Pb and its compounds	189	37	177
Hg and its compounds	1.1	0.4	250
Chlorfenvinphos (2-Chloro-1-(2,4-dichlor...	1.1	0.36	22
Ni and its compounds	85	47	70
Benzene	147	59	74
Hexachlorobutadiene	12	0.39	41
Diuron (3-(3,4-dichlorophenyl)-1,1-dimet...	13	7.1	146
Dichloromethane	486	320	20
Alachlor (2-Chloro-N-(2,6-diethylphenyl)...	4.2	4	38
Naphthalene	30	7.8	29
Fluoranthene	0.96	0.098	47
Trichloromethane	72	44	137
Trichlorobenzene	2.4	2	25
Hexachlorobenzene	16	7.6	25
1,2-Dichloroethane	401	335	31
Di(2-ethylhexyl)phthalate	62	0.32	56
4-(1,1',3,3'-tetramethylbutyl)-phenol	0.91	0.41	4
Pentachlorobenzene	0.35	0.3	9
Anthracene and its compounds	4.6	0.2	4

Load Phylogenetic Tree

The phylogenetic tree for the fish species together with branch lengths was obtained from timetree.org. Species that were missing in the timetree database were excluded from the plot as phylogenetically informed regressions can't be done without phylogenetic distances.